## 1. Introduction

The world has entered the digital world. Various technologies, including the computer, facilitate various human activities. Various organizations, groups, and even individuals compete to get the latest and best information. Often the method used to obtain this information could be more effective. To make information processing more effective, one of the steps taken is using machine learning. However, sometimes the information obtained could be in its better form, and sometimes the data obtained still needs to be more efficient; this is often found in large data sets. These large data sets usually contain imbalanced classes, which may affect classification performance[1].

Data sets that contain significant differences between classes with very few instances, known as minor classes [2], and classes with sufficient instances, known as major classes, are considered imbalanced data sets. Synthetic Minority Oversampling Technique (SMOTE) is a frequently used oversampling technique to handle imbalance class problems and is also deemed a very successful technique to generate synthetic data [3]. A simple explanation of how SMOTE works. This technique starts with selecting an instance from the minority class to be selected as a point, identifying its K-nearest neighbors, then creating a line to its K-nearest neighbor. Points between this line are what is considered synthetic observation. As SMOTE generates synthetic observation on all the minority class data, this affects the original data distribution of the minority class. As an alternative to mitigate such a problem, ADASYN is proposed as a possible solution [4]. Adasyn generates synthetic observations based on a weighing technique based on more complex data points of the minority class that are harder to classify.

The classification method we are using is Random Forest Classifier. In order to create an efficient classification model, our research includes data resampling with SMOTE and ADASYN. The purpose of this research is to compare the effect of imbalanced data sets on classification performance and how an effective oversampling method can improve the performance of the model that has been built.

In this research, topics, and limitations are applied to find the performance of the model by applying several research scenarios. The limitation of this research is that classification using Random Forest (RF) is carried out on imbalanced data, then applying SMOTE / ADASYN to return to RF classification. The data used is the e-coli data set obtained from the KEEL website. The data taken has an imbalanced ratio range of nine to thirteen. In this study, there are two objectives, namely, to see how the performance and accuracy of the model in analyzing imbalanced data using the SMOTE oversampling method and to compare with the ADASYN oversampling method.

The handling imbalanced data set with SMOTE and ADASYN tries to solve the imbalanced classification problems. We applied the oversampling technique to improve the class situation on the E coli data set. Previous research rarely compares the raw performance of the oversampling technique, especially on the E coli data set. We show how SMOTE and ADASYN work and evaluate how each sampling method affects the classification performance. In our research, we use the balanced accuracy score as a comparable metric that covers the whole performance of our classifier. This research aims to get a comparable result on which sampling methods work best and does every data set need to be resampled to get the best classification performance.